Application of the Information Criterion to the Estimation of Galaxy Luminosity Function

Tsutomu T. Takeuchi *
(e-mail: takeuchi@kusastro.kyoto-u.ac.jp)
Department of Astronomy, Faculty of Science, Kyoto University, 606-8502, Japan

Abstract. To determine the exact shape of the luminosity function (LF) of galaxies is one of the central problems in galactic astronomy and observational cosmology. The most popular method to estimate the LF is maximum likelihood, which is clearly understood with the concepts of the information theory. In the field of information theory and statistical inference, great advance has been made by the discovery of Akaike's Information Criterion (AIC). It enables us to perform a direct comparison among different types of models with different numbers of parameters, and provides us a common basis of the model adequacy. In this paper we applied AIC to the determination of the shape of the LF. We first treated the estimation using stepwise LF (Efstathiou, Ellis, & Peterson, 1988), and derived a formula to obtain the optimal bin number. In addition, we studied the method to compare the goodness-of-fit of the parametric form (Sandage, Tamman, & Yahil, 1979) with stepwise LF.

Keywords: Cosmology — galaxies: luminosity function – methods: statistical

1. Introduction

The luminosity function of galaxies (LF) is one of the fundamental descriptions of the galaxy population (e.g. Bingelli, Sandage, & Tammann, 1988). It is also essential to interpret the galaxy number counts (e.g. Koo & Kron, 1992; Ellis, 1997) or to analyse galaxy clustering (e.g. Strauss & Willick, 1995; Efstathiou, 1996). Furthermore, the LF is a fundamental test for the theory of galaxy formation (e.g. Baugh, Cole, & Frenk, 1996). Recently, the exact shape of the LF has been of particular interest, because it is one of the key issues to the "faint blue galaxy problem" of galaxy number counts (Koo & Kron, 1992; Ellis, 1997), and may be related to dwarf galaxy formation (e.g. Babul & Rees, 1992; Babul & Ferguson, 1996 ; Hogg & Phinney, 1997; Ferguson & Babul, 1998).

Instead of classical V/V_{m}-estimator of the LF (Schmidt, 1968), the maximum likelihood method, which is free of bias induced by density inhomogeneity, is popular among recent studies. Various techniques have been proposed by expert astronomers (Lynden-Bell, 1971; Marshall et al., 1983; Chołoniewski, 1986; Caditz & Petrosian, 1993). The

* Research Fellow of the Japan Society for the Promotion of Science

parametric method of Sandage, Tamman, & Yahil (1979) (STY) and the stepwise maximum likelihood method by Efstathiou, Ellis, & Peterson (1988)(EEP) are the most popular among them. And in addition, Lynden-Bell's method is quite sophisticated and requires no assumption for the probability density function. But in practice, we often need to smooth the obtained LF, because the discrete feature of the real data is not suitable for various studies. When we perform smoothing or binning, it remains unclear, for example, that how many numbers of bins we should take. Too wide bin leads to underestimation of the slope, and too narrow bin makes the results unstable (e.g. Caditz & Petrosian, 1993; Strauss & Willick, 1995; Heyl et al., 1997). It is also hard to estimate the relative goodness between different types of statistical models. We should face this kind of problem, for example, when we try to examine whether the Schechter form (Schechter, 1976) provides an acceptable fit or not, using the likelihood ratio of the models (EEP). In general, the more the number of free parameter is, the better the fitting becomes, and consequently, *the likelihood function gets larger*.

In order to evaluate the goodness-of-fit of a certain model to the data, the concept of the information criterion is useful. Since the middle of 1970's, vast advances have been made in the field of the statistical inference by the discovery of Akaike's Information Criterion (AIC: Akaike, 1974). The meaning of the AIC is clearly understood as an extention of the maximum likelihood method, and closely related to the information entropy, especially to the 'relative entropy' of two probability distributions. The relative entropy has a property just like a distance in differential geometry, i.e. it is a distance between the two probability distributions. Using AIC enables us to compare the goodness of a certain model with that of another type directly. For this fascinating property, AIC and its cousins are applied to the various fields of studies concerning statistical model selection.

In this paper, we make an attempt to apply AIC to the estimation problem of the LF. In order to understand the Akaike's theory, some knowledges from information theory are required. Section 2 is devoted to this mathematical background concepts. In section 3, first we see how AIC is applied to decide the step number of EEP method, and next how to judge the goodness of fit of the Schechter form estimated by STY method. Our summary is presented in section 4.

2. Akaike's information criterion

In this section, we make an informal introduction of Akaike's theory. We do not try to be mathematically rigorous, but make an attempt to make it comprehensible.

2.1. Kullback–Leibler information and information matrix

First of all, we consider the 'information entropy'. Here we consider the (discrete) probability distribution $\{f_i\}_{i=1,\cdots,N}$. The self information is defined as

$$I(f_i) = -\ln f_i \ , \tag{1}$$

where $\ln f_i = \log_e f_i$. Then, the expectation value of $I(f_i)$ is

$$S \equiv \mathsf{E}\,[I] = -\sum_{i=1}^{N} f_i \ln f_i \ . \tag{2}$$

This is called the information entropy. When we have two different probability distributions $\{f_i\}_{i=1,\cdots,N}$ and $\{g_i\}_{i=1,\cdots,N}$, we can construct the following quantity:

$$V(f,g) \equiv -\sum_{i=1}^{N} f_i \ln \frac{f_i}{g_i} \ . \tag{3}$$

This is called Kullback–Leibler information (Kullback & Leibler, 1951), or more comprehensively, relative entropy of $\{f_i\}_{i=1,\cdots,N}$ and $\{g_i\}_{i=1,\cdots,N}$. For understanding the meaning of $V(f,g)$, we define $p_i = p_i(\theta_1,\cdots,\theta_K)$ $(i = 1,\cdots,N)$, where $(\theta_1,\cdots,\theta_K)$ are the parameters on which the model depends, and let

$$\begin{cases} f_i = p_i(\theta_1^0,\cdots,\theta_K^0) \ , \\ g_i = p_i(\theta_1^0 + \mathrm{d}\theta_1,\cdots,\theta_K^0 + \mathrm{d}\theta_K) \ . \end{cases} \tag{4}$$

Then, after some arithmetics, we have

$$V(f,g) = \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K} I(\theta^0)_{kl}\mathrm{d}\theta_k\mathrm{d}\theta_l \ , \tag{5}$$

where

$$I_{kl}(\theta^0) = \sum_{i=1}^{N} p_i \left(\frac{\partial \ln p_i}{\partial \theta_k}\frac{\partial \ln p_i}{\partial \theta_l}\right)_{\theta=\theta^0} \ . \tag{6}$$

The subscript $\theta = \theta^0$ means $\theta_i = \theta_i^0$ for $i = 1, \cdots, N$. The matrix I_{kl} is called Fisher's information matrix. If we regard I_{kl} as a "metric" of the parameter space of θ_k and θ_l, we can treat $V(f, g)$ as the distance between the two probability distributions $\{f_i\}_{i=1,\cdots,N}$ and $\{g_i\}_{i=1,\cdots,N}$, just as we do in differential geometry.

2.2. Maximum likelihood estimation in the context of information theory

We, here, will see the maximum likelihood method considering relative entropy $V(f, g)$. Maximum likelihood is the method in order to estimate the optimal model for *a given set of data*. Let $x_1, \cdots, x_N$ as a realization of random variable X which obeys the unknown probability distribution $f(X)$. We define a model probability distribution $g(x; \theta_1, \cdots, \theta_K)$, which depends on K parameters $\{\theta_k\}_{k=1,\cdots,K}$. Then, the likelihood function $\mathcal{L}$ is defined by

$$\mathcal{L} \equiv \mathcal{L}(\theta_1, \cdots, \theta_K | x_1, \cdots, x_n) = \prod_{n=1}^{N} g(x_n; \theta) \ , \tag{7}$$

and the logarithmic likelihood is

$$\ln \mathcal{L} \equiv \sum_{n=1}^{N} \ln g(x_n; \theta) \ . \tag{8}$$

Performing maximum likelihood estimation is to find the parameter set $\theta = \theta_1, \cdots, \theta_K$ which defines the most suitable model $g(x; \theta)$ for unknown *true* probability distribution $f(X)$. If we know the true $f(X)$, using relative entropy eq.(3), we have

$$\begin{aligned} V(f, g) &= \sum_{n=1}^{N} f(x_n) \ln \frac{f(x_n)}{g(x_n; \theta)} = \mathsf{E}\left[\ln \frac{f(X)}{g(X; \theta)}\right] \ , \\ &= \mathsf{E}\left[\ln f(X)\right] - \mathsf{E}\left[\ln g(X; \theta)\right] \ . \end{aligned} \tag{9}$$

But actually we do not know $f(X)$, we cannot obtain $\mathsf{E}[\,\cdot\,]$. Therefore we use the law of large numbers.

THEOREM 1. **(The law of large numbers)** *Let $X_1, \cdots, X_N$ independent random variables, which have the same mean value*

$$\mathsf{E}[X_1] = \cdots = \mathsf{E}[X_N] = m \ , \tag{10}$$

and $^\exists\sigma^2 > 0$, such that

$$\mathsf{E}[(X_i - m_i)] \leq \sigma^2 \quad \text{for} \quad i = 1, \cdots, N \ . \tag{11}$$

Then for a random variable

$$X \equiv \frac{1}{N}\sum_{i=1}^{N} X_i \ , \tag{12}$$

we have

$$\mathsf{E}\left[X\right] = m \ , \tag{13}$$

as $N \to \infty$.

Proof of this famous theorem is seen in many textbooks on probability theory, so we omit it (see e.g. Shiryaev, 1996).

Hence, if the size of the dataset N is sufficiently large,

$$\frac{1}{N}\ln\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\ln g(x_n;\theta) \simeq \mathsf{E}\left[\ln g(X;\theta)\right] \ . \tag{14}$$

It leads to

$$V(f,g) = \mathsf{E}\left[\ln f(X)\right] - \frac{1}{N}\sum_{i=1}^{N}\ln g(x_n;\theta) \ . \tag{15}$$

The last term can be obtained without knowing true $f(X)$. Since $\mathsf{E}\left[\ln f(X)\right]$ is independent of $\theta = \theta_1, \cdots, \theta_K$, what to be maximized is $\ln\mathcal{L} = \sum_{i=1}^{N}\ln g(x_n;\theta)$. Consequently, we obtain the likelihood equation:

$$\left.\frac{\partial}{\partial\theta_k}\ln\mathcal{L}\right|_{\theta=\hat{\theta}} = \frac{\partial}{\partial\theta_k}\left\{\sum_{n=1}^{N}\ln g(x_n;\theta)\right\}_{\theta=\hat{\theta}} = 0 \tag{16}$$

for $k = 1, \cdots, K$, where we defined $\hat{\theta}$ to satisfy

$$\max_{\theta\in\Theta}(\ln\mathcal{L}) = \ln\mathcal{L}(\hat{\theta}|x_1, \cdots, x_n) \ . \tag{17}$$

Here, Θ denotes the family of parameters θ. The solution $\hat{\theta}$ of the eq. (16) is called the maximum likelihood estimator.

We, next, consider the error of the logarithmic likelihood equation (16):

$$\Delta(\hat{\theta}) \equiv \left.\mathsf{E}\left[\ln g(X;\theta)\right]\right|_{\theta=\hat{\theta}} - \frac{1}{N}\sum_{n=1}^{N}\left.\ln g(x_n;\theta)\right|_{\theta=\hat{\theta}} \ . \tag{18}$$

Let $\theta^0 = \theta_1^0, \cdots, \theta_K^0 \in \Theta$ the estimator which maximize $\mathsf{E}\left[\ln g(X;\theta)\right]$. Taylor expansion and theorem 1 lead the well-known important result:

$$\Delta(\hat{\theta}) \ \simeq \ \sum_{k=1}^{N}\sum_{l=1}^{N}(\hat{\theta}_k - \theta_k^0)\left.\mathsf{E}\left[\frac{\partial^2}{\partial\theta_k\partial\theta_l}\ln g(X;\theta)\right]\right|_{\theta=\theta^0}(\hat{\theta}_l - \theta_l^0)$$

$$\simeq -\sum_{k=1}^{N}\sum_{l=1}^{N}(\hat{\theta}_k - \theta_k^0) I_{kl} (\hat{\theta}_l - \theta_l^0) \; . \tag{19}$$

Again, I_{kl} is the information matrix. Thus, we can evaluate the error of likelihood function using I_{kl} (e.g. Stuart, Ord, & Arnold, 1999).

2.3. Akaike's Information Criterion (AIC)

In this subsection, we discuss Akaike's information criterion (AIC). To the first order of θ, likelihood equation (16) is expressed as

$$\sum_{n=1}^{N}\left\{\frac{\partial}{\partial\theta_k}\ln g(x_n;\theta)\right\}_{\theta=\theta^0} + \sum_{n=1}^{N}\sum_{l=1}^{K}\left\{\frac{\partial^2}{\partial\theta_k\partial\theta_l}\ln g(x_n;\theta)\right\}_{\theta=\theta^0}(\hat{\theta}_l - \theta_l^0) \\ = 0 \; . \tag{20}$$

In order to evaluate this equation, we use the multivariate central limit theorem.

THEOREM 2. **(Multivariate central limit theorem)** *Let a_n^k ($n = 1, \cdots, N$) a set of sample values of a vector random variable A^k, which have a mean*

$$\mathsf{E}_A\left[A^k\right] = m^k \; , \tag{21}$$

($m^k < \infty$) and a dispersion

$$\mathsf{E}_A\left[(A^k - m^k)(A^l - m^l)\right] = \sigma^{kl} \tag{22}$$

for ($k = 1, \cdots, K$). We define Z^k as a vector random variable whose sample value z^k is

$$z^k \equiv \sqrt{N}\left\{\frac{1}{N}\sum_{n=1}^{N} a_n^k - m^k\right\} \; . \tag{23}$$

Then, as $N \to \infty$, Z^k becomes a Gaussian vector random variable with

$$\mathsf{E}_Z\left[Z^k\right] = 0 \; , \quad \mathsf{E}_Z\left[Z^k Z^l\right] = \sigma^{kl} \; . \tag{24}$$

The proof of this theorem is also incorporated in many textbooks (e.g. Shiryaev 1996).

By definition of θ^0, we have

$$\mathsf{E}\left[\frac{\partial}{\partial\theta_k}\ln g(X;\theta)\right]\bigg|_{\theta=\theta^0} = \left\{\frac{\partial}{\partial\theta_k}\mathsf{E}\left[\ln g(X;\theta)\right]\right\}_{\theta=\theta^0} \equiv \mathcal{M}^k = 0 \; . \tag{25}$$

Therefore, with eq. (6)

$$\begin{aligned}
&\mathsf{E}\left[\left(\frac{\partial}{\partial\theta_k}\ln g(X;\theta)-\mathcal{M}^k\right)\left(\frac{\partial}{\partial\theta_l}\ln g(X;\theta)-\mathcal{M}^l\right)\right]\Bigg|_{\theta=\theta^0}\\
&=\mathsf{E}\left[\left(\frac{\partial}{\partial\theta_k}\ln g(X;\theta)\right)\left(\frac{\partial}{\partial\theta_l}\ln g(X;\theta)\right)\right]\Bigg|_{\theta=\theta^0}\\
&=I_{kl}\ . \qquad (26)
\end{aligned}$$

Using the result of theorem 2, we have a vector random variable Z^k which is a Gaussian with

$$\mathsf{E}_Z[Z^k]=0\ ,\ \mathsf{E}_Z[Z^kZ^l]=I_{kl}\ , \qquad (27)$$

and whose sample value z^k is

$$z^k=\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\left\{\frac{\partial}{\partial\theta_k}\ln g(x_n;\theta)\right\}_{\theta=\theta^0}\ . \qquad (28)$$

We, then, apply theorem 1 to the second term of the eq.(20), it follows that

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\partial^2}{\partial\theta_k\partial\theta_l}\ln g(x_n;\theta)\right\}_{\theta=\theta^0} &\simeq \mathsf{E}\left[\frac{\partial^2}{\partial\theta_k\partial\theta_l}\ln g(X;\theta)\right]\Bigg|_{\theta=\theta^0}\\
&= -I_{kl}\,. \qquad (29)
\end{aligned}$$

Substituting the eqs. (28) and (29) into eq. (20) gives

$$z^k=\sqrt{N}\sum_{l=1}^{K}I_{kl}(\hat\theta_l-\theta_l^0)\ . \qquad (30)$$

If I_{kl} is non-singular (We can always expect it by reparametrization), the inverse matrix I_{kl}^{-1} is defined as

$$\sum_{k=1}^{K}I_{uk}I_{kv}^{-1}=\delta_{uv} \qquad (31)$$

where δ_{uv} is a Kronecker's delta, and it leads to

$$(\hat\theta_k-\theta_k^0)=\frac{1}{\sqrt{N}}\sum_{l=1}^{K}I_{kl}^{-1}z^l\ . \qquad (32)$$

Hence, we can write

$$\Delta(\hat\theta)=-\sum_{k=1}^{K}\sum_{l=1}^{K}(\hat\theta_k-\theta_k^0)I_{kl}(\hat\theta_l-\theta_l^0)=-\frac{1}{N}\sum_{u=1}^{K}\sum_{v=1}^{K}I_{uv}^{-1}z^uz^v\ . \qquad (33)$$

Now, again we use theorem 2, and eventually we obtain,

$$\begin{aligned}\Delta(\hat{\theta}) &\simeq -\frac{1}{N}\sum_{u=1}^{K}\sum_{v=1}^{K} I_{uv}^{-1}\mathsf{E}_Z[Z^u Z^v] = -\frac{1}{N}\sum_{u=1}^{K}\sum_{v=1}^{K} I_{uv}^{-1} I_{uv} \\ &= -\frac{K}{N} \,. \end{aligned} \tag{34}$$

Thus, we reach final important result:

$$\begin{aligned}\mathsf{E}[\ln g(X;\theta)]\big|_{\theta=\hat{\theta}} &\simeq \frac{1}{N}\sum_{n=1}^{N}\ln g(x_n;\theta)\bigg|_{\theta=\hat{\theta}} - \frac{K}{N} \\ &= \frac{1}{N}\Big(\ln\mathcal{L}(\hat{\theta}|x_1,\cdots,x_n) - K\Big) \,. \end{aligned} \tag{35}$$

Since, as eq. (9) shows, the distance between the true probability distribution $f(X)$ and model $g(X;\theta)$ is evaluated by $\mathsf{E}[\ln g(X;\theta)]$, we have to choose $\hat{\theta} \in \Theta$ to be maximize the quantity $(\ln\mathcal{L} - K)$ for given K. Though large K provides us large $\ln\mathcal{L}$, it also reduces $(\ln\mathcal{L} - K)$ at the same time, therefore the optimal number of parameters K has to be taken so that it maximize $(\ln\mathcal{L}(\hat{\theta}) - K)$. For historical reason (e.g. Rao, 1965), the value $-2(\ln\mathcal{L}(\hat{\theta}) - K)$ has been considered (see section 4). It is the Akaike's information criterion (AIC). We, now, see that the optimal model is the one which minimizes AIC.

3. Application of AIC to LF estimation

3.1. The stepwise maximum likelihood method and optimal step number

We apply AIC to the estimation of the shape of the LF. First we consider the stepwise maximum likelihood method introduced by EEP. Throughout this paper, for simplicity, we do not deal with some issues which turn out to be important when we treat the real redshift data, e.g. errors in magnitude measurement, surface brightness incompleteness, etc. These practical problems are summarized and dealt with in e.g. Lin et al. (1996).

The EEP method uses the form of the LF

$$\phi(M) = \sum_{k=1}^{K}\phi_k W(M_k - M) \tag{36}$$

where M is the absolute magnitude of a galaxy, which is obtained by

$$M = m - 25 - 5\log d_L(z) - k(z) \,. \tag{37}$$

Here $\log \equiv \log_{10}$, m : the apparent magnitude, $d_L(z)$ [Mpc] : the luminosity distance corresponding to redshift z, and $k(z)$ is the K-correction. The window function $W(M_l - M)$ is defined by

$$W(M_l - M) \equiv \begin{cases} 1 & M_l - \Delta M/2 \le M \le M_l + \Delta M/2 \ , \\ \\ 0 & \text{otherwise} \ . \end{cases} \tag{38}$$

In terms of information theory, the stepwise LF is regarded as a histogram model, a kind of discrete probability distribution models. The parameters of the model $\theta \in \Theta$ is in this case $\theta_k = \phi_k$ $(k = 1, \cdots, K)$ themselves. Given the galaxy redshift survey data of size N, we set

$$\Delta M = \frac{M_{\rm upper} - M_{\rm lower}}{K - 1} \ , \tag{39}$$

$$M_{\rm upper} \equiv \max_{i=1,\cdots,N} \{M_i\} \ , \ M_{\rm lower} \equiv \min_{i=1,\cdots,N} \{M_i\} \ .$$

The denominator of the eq. (39) is $K - 1$, though the number of bins are K, because M_k is evaluated at the center of k-th bin, and therefore the magnitude range becomes $M_{\rm lower} - \Delta M/2 \sim M_{\rm upper} + \Delta M/2$. A certain constraint is adopted to $\{\phi_k\}_{k=1,\cdots,K}$ in usual manner. However, since the normalization does not used in the determination of the shape of the LF (otherwise a Lagrange multiplier λ would appear in the denominator of the eq. (44)), and neither adopted in the procedure of the following STY method, we do not set any constraint in our formulation. According to EEP, the likelihood function is

$$\begin{aligned} &\mathcal{L}(\{\phi_k\}_{k=1,\cdots,K} | \{M_i\}_{i=1,\cdots,N}) \\ &= \prod_{i=1}^{N} \frac{\sum_{l=1}^{K} W(M_l - M_i)\phi_l}{\sum_{l=1}^{K} \phi_l H(M_{\rm lim}(z_i) - M_l)\Delta M} \ , \end{aligned} \tag{40}$$

$$\begin{aligned} &H(M_{\rm lim}(z_i) - M_l) \\ &\equiv \begin{cases} 1 & M_{\rm lim}(z_i) - \Delta M/2 > M_l \\ \\ \dfrac{M_{\rm lim}(z_i) - M_l}{\Delta M} + \dfrac{1}{2} & M_{\rm lim}(z_i) - \Delta M/2 \le \\ & \quad M_l < M_{\rm lim}(z_i) + \Delta M/2 \\ 0 & M_{\rm lim}(z_i) + \Delta M/2 \le M_l \end{cases} \end{aligned} \tag{41}$$

where $M_{\rm lim}(z_i)$ is the absolute magnitude corresponding to the survey limit $m_{\rm lim}$ at redshift z_i. This likelihood function $\mathcal{L}$ clearly depends on the bin width ΔM, and consequently, its likelihood ratio to other

model depends on ΔM. This has been regarded as an "artificial effect" to be eliminated by certain procedures, but it is not true because the choice of ΔM is in this case the selection of histogram model itself.

The logarithmic likelihood is expressed as

$$\ln \mathcal{L} = \sum_{i=1}^{N} \left[\sum_{l=1}^{K} W(M_l - M_i) \ln \phi_l \right.$$
$$\left. - \ln \left\{ \sum_{l=1}^{K} \phi_l H(M_{\text{lim}}(z_i) - M_l) \Delta M \right\} \right] . \tag{42}$$

Hence, likelihood equation becomes

$$\frac{\partial \ln \mathcal{L}}{\partial \phi_k} = \sum_{i=1}^{N} \frac{W(M_k - M_i)}{\phi_k}$$
$$- \sum_{i=1}^{N} \frac{H(M_{\text{lim}}(z_i) - M_k)\Delta M}{\sum_{l=1}^{K} \phi_l H(M_{\text{lim}}(z_i) - M_l)\Delta M} = 0 \tag{43}$$

and it reduces to

$$\phi_k \Delta M = \frac{\sum_{i=1}^{N} W(M_k - M_i)}{\sum_{i=1}^{N} \dfrac{H(M_{\text{lim}}(z_i) - M_k)}{\sum_{l=1}^{K} \phi_l H(M_{\text{lim}}(z_i) - M_l)\Delta M}} . \tag{44}$$

This equation can be solved by iteration, and we obtain the maximum likelihood estimator $\hat{\phi} = \{\hat{\phi}_k\}_{k=1,\cdots,K}$. Thus,

$$\text{AIC}_{\text{EEP}} = -2(\ln \mathcal{L}|_{\phi=\hat{\phi}} - K) . \tag{45}$$

Therefore, the step number K should be taken so that it minimizes the eq. (45).

REMARK 1. *The obtained number K may not stand for the number of physical parameters, viz. when we get a certain K, it does not mean we need K physical quantities for explanation. The obtained stepwise LF is the one which best reflects the property of the underlying data population.*

3.2. Comparison of Schechter form with stepwise LF

We, here, consider how to compare the goodness of fit of the Schechter form to that of EEP stepwise LF. We set the LF as

$$\phi(M) = 0.4 \ln 10 \; \phi^* \left(10^{0.4(M_* - M)}\right)^{1+\alpha} \exp\left(-10^{0.4(M_* - M)}\right) . \tag{46}$$

The likelihood function is therefore

$$\mathcal{L}(\{\alpha, M_*\}|\{M_i\}_{i=1,\cdots,N})$$
$$= \prod_{i=1}^{N} \frac{\left(10^{0.4(M_*-M_i)}\right)^{1+\alpha} \exp\left(-10^{0.4(M_*-M_i)}\right)}{\int_{-\infty}^{M_{\rm lim}(z_i)} \left(10^{0.4(M_*-M)}\right)^{1+\alpha} \exp\left(-10^{0.4(M_*-M)}\right) \mathrm{d}M} . \quad (47)$$

Thus the logarithmic likelihood becomes

$$\begin{aligned}
\ln \mathcal{L} \; &= \; 0.4 \ln 10 \,(1+\alpha) \sum_{i=1}^{N} (M_* - M_i) - \sum_{i=1}^{N} 10^{0.4(M_*-M_i)} \\
&\quad - \sum_{i=1}^{N} \ln \int_{-\infty}^{M_{\rm lim}(z_i)} \left(10^{0.4(M_*-M)}\right)^{1+\alpha} \exp\left(-10^{0.4(M_*-M)}\right) \mathrm{d}M \\
&= \; 0.4 \ln 10 \,(1+\alpha) \sum_{i=1}^{N} (M_* - M_i) \\
&\quad - \sum_{i=1}^{N} 10^{0.4(M_*-M_i)} - 0.4 \ln 10 \sum_{i=1}^{N} \ln \Gamma(1+\alpha, y_i) \; , \qquad (48)
\end{aligned}$$

where

$$y_i \equiv 10^{0.4(M_*-M_{\rm lim}(z_i))} \; , \qquad (49)$$
$$\Gamma(z,p) \equiv \int_{p}^{\infty} t^{z-1} e^{-t} \mathrm{d}t \; . \qquad (50)$$

Equation (50) is known as Legendre's incomplete gamma function.

Since the parameters to be estimated are α and M_* in the STY method, we have the set of likelihood equations:

$$\begin{aligned}
&\frac{\partial \ln \mathcal{L}}{\partial \alpha} \\
&= 0.4 \ln 10 \left\{ \sum_{i=1}^{N} (M_* - M_i) - \sum_{i=1}^{N} \frac{\int_{y_i}^{\infty} \ln t \; t^{\alpha} e^{-t} \mathrm{d}t}{\Gamma(1+\alpha, y_i)} \right\} = 0 \; , \qquad (51) \\
&\frac{\partial \ln \mathcal{L}}{\partial M_*} = 0.4 \ln 10 (1+\alpha) N \\
&-0.4 \ln 10 \sum_{i=1}^{N} 10^{0.4(M_*-M_i)} - 0.4 \ln 10 \sum_{i=1}^{N} \frac{{y_i}^{\alpha} e^{-y_i}}{\Gamma(1+\alpha, y_i)} \frac{\partial y_i}{\partial M_*} \\
&= 0.4 \ln 10
\end{aligned}$$

$$\times \left\{ (1+\alpha)N - \sum_{i=1}^{N} 10^{0.4(M_* - M_i)} - 0.4 \ln 10 \frac{{y_i}^{1+\alpha} e^{-y_i}}{\Gamma(1+\alpha, y_i)} \right\}$$
$$= 0 \; . \qquad (52)$$

By solving the eqs. (51) and (52), we obtain the maximum likelihood estimators $\hat{\alpha}$ and $\hat{M}_*$. We have two free parameters in the STY method, and we consequently have the AIC of STY method

$$\mathrm{AIC}_{\mathrm{STY}} = -2(\ln \mathcal{L}_{\mathrm{STY}}|_{\theta=\hat{\theta}} - 2) \; , \qquad (53)$$

where the subscript $\theta = \hat{\theta}$ represents that $\alpha = \hat{\alpha}$ and $M_* = \hat{M}_*$.

The relative goodness of fit of the Schechter LF compared with the stepwise one is evaluated by

$$\begin{aligned} \Delta\mathrm{AIC} &\equiv \mathrm{AIC}_{\mathrm{STY}} - \mathrm{AIC}_{\mathrm{EEP}} \\ &= -2(\ln \mathcal{L}_{\mathrm{STY}}|_{\theta=\hat{\theta}} - \ln \mathcal{L}_{\mathrm{EEP}}|_{\phi=\hat{\phi}} + K - 2) \; . \qquad (54) \end{aligned}$$

4. Summary and discussion

In in the previous sections we introduced the Akaike's information criterion (AIC) to the maximum likelihood estimation of galaxy luminosity function (LF). The AIC is closely related to the "distance" between two probability distributions which becomes clear by using Fisher's information matrix. It is expressed as

$$\mathrm{AIC} = -2(\ln \mathcal{L}(\hat{\theta}) - K) \; ,$$

where $\mathcal{L}$ is a likelihood function, $\hat{\theta}$ is a set of maximum likelihood estimators, and K is the number of free parameters of the assumed model. Since the concept of the information criterion seems unfamiliar to the astronomical community, we discuss its meaning and practical use in this section.

What we must stress is that the difference between the χ^2-type goodness-of-fit quantities and AIC. The statistic usually used to evaluate uncertainty of the LF estimation is the χ^2, which was extensively discussed by EEP. This error evaluation is based on the well-known fact that the logarithmic likelinood ratio, $-2\ln(\mathcal{L}(\theta)/\mathcal{L}(\theta^0))$, is asymptotically distributed as χ^2 distribution (e.g. Rao, 1965; Stuart, Ord, & Arnold, 1999). The likelihood ratio is regarded as a random variable, and discussed with respect to its confidence level. In order to estimate its distribution, EEP performed a Monte Carlo simulations and confirmed their error estimation. On the other hand, as we discussed in

section 2, the information criterion is a value obtained as a result of limit theorems, and is not regarded as a random variable. In statistical model selection, we often use such kind of goodness-of-fit index. The class of information criteria including AIC was invented along with such concepts. Thus, though the AIC is related to the χ^2 statistic, likelihood ratio, its value is not discussed with a confidence level (see Sakamoto, Ishiguro, & Kitagawa, 1986 for details).

Then, how should we treat the AIC value for practical use? In the case of stepwise LF model (EEP), K is a number of steps of the LF. The AIC is

$$\mathrm{AIC}_{\mathrm{EEP}} = -2(\ln \mathcal{L}|_{\phi=\hat{\phi}} - K) \ ,$$

where $\phi = \phi_1, \cdots, \phi_K$ is the step heights. We should compare the $\mathrm{AIC}_{\mathrm{EEP}}$ and choose the number K which minimizes $\mathrm{AIC}_{\mathrm{EEP}}$. The larger the number of parameter is, the larger datasize N is required, because each parameter estimation procedure carries its own error. In case K is considerably larger in comparison with the data size N, we cannot use the limit theorems of probability theory like theorem 1, since the results would no longer hold. Thus, number of parameters should be taken as

$$K \lesssim 2\sqrt{N} \tag{55}$$

and at most

$$K < \frac{N}{2} \ . \tag{56}$$

Bin number K comparable to N is meaninglessly large because such a fine binning yields horribly numerous empty bins. In such case, the AIC diverges by its definition, which means that the bin number is not a good choice. Sometimes, AIC becomes smaller and smaller as K is larger, and does not take minimum value. Then the assumed form of the model is significantly wrong, so that $V(f, g)$ is very large.

We next consider the comparison between the stepwise form and parametric models. For the STY method, free parameters are α and M_* in the Schechter function (eq.(46)). In order to compare the goodness of fit of EEP and STY models, we can use the difference of the AIC

$$\begin{aligned} \Delta\mathrm{AIC} &\equiv \mathrm{AIC}_{\mathrm{STY}} - \mathrm{AIC}_{\mathrm{EEP}} \\ &= -2(\ln \mathcal{L}_{\mathrm{STY}}|_{\theta=\hat{\theta}} - \ln \mathcal{L}_{\mathrm{EEP}}|_{\phi=\hat{\phi}} + K - 2) \ , \end{aligned}$$

where the $\theta = \hat{\theta}$ means that $\alpha = \hat{\alpha}$ and $M_* = \hat{M}_*$. As we mentioned above, the evaluation of AIC is essentially different from that of χ^2 statistic. Originally, EEP regarded their stepwise LF as a *true* probability density function and derived the likelihood ratio. They mentioned

that the likelihood ratio obtained from stepwise and Schechter LF was very large and negative. This fact is a natural consequence, because, from the aspect of (relative) Kullback–Leibler distance, stepwise LF has much larger parameters compared with parametric Schechter LF and therefore the goodness-of-fit is much better. Information criterion is suitable for such problem.

Again we note that it is the difference of AIC values that matters and not the absolute values themselves. This is because we would never know the ture distribution from a finite size data, and we used the sample expectation instead of the expectation value based on true distribution (see eq. (15)). In our point of view, the stepwise LF is an estimates derived along with one of the family of statistical models in this paper. Thus the difference of AICs, ΔAIC can be used for our purpose. Then, what should we regard as the "scale" of ΔAIC ? The order of the variation of the AIC is that of the number of parameters, K, which is an integer. Thus, when we have the difference of AICs

$$\Delta\mathrm{AIC} \gtrsim 1 \ , \tag{57}$$

the two distributions are significantly different. Generally, $\Delta\mathrm{AIC} = \mathrm{AIC}_{\mathrm{STY}} - \mathrm{AIC}_{\mathrm{EEP}}$ is larger than unity, therefore we judge that the goodness-of-fit is not sufficient, and other functional form can instead be used. But acutually, for the optical galaxy LF, we are interested in comparison between the goodness-of-fit of Schechter form and other forms. The functional form choice is also an interesting issue in the estimaiton of the *IRAS* galaxy LF, which is known to be significantly differnt from Schechter form (e.g. Saunders et al., 1996). In such case AIC works as a powerful tool for model selection of the fitting functions.

Acknowledgements

We offer our gratitude to Drs. Hiroyuki Hirashita and Kohji Yoshikawa who have made critical reading of the original version of the paper. We also thank Drs. Toru Yamada, Takashi Ichikawa, Hajime Sugai, and Takehiko Wada for useful discussions and comments. We acknowledge the Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

References

Akaike, H. 1974, *IEEE Trans. Autom. Contr.*, **AC-19**, 716

Babul, A., & Ferguson, H. C. 1996, *Astrophys. J*, **458**, 100
Babul, A., & Rees, M. J. 1992, *Mon. Not. R. Astron. Soc.*, **255**, 346
Baugh, G. M., Cole, S., & Frenk, C. S. 1996, *Mon. Not. R. Astron. Soc.*, **283**, 1361
Bingelli, B., Sandage, A., & Tammann, G. A. 1988, *Ann. Rev. Astron. Astrophys.*, **26**, 509
Caditz, D., & Petrosian, V. 1993, *Astrophys. J*, **416**, 450
Chołoniewski, J. 1986, *Mon. Not. R. Astron. Soc.*, **223**, 1
Efstathiou, G., Ellis, R. S., & Peterson, B. A. 1988, *Mon. Not. R. Astron. Soc.*, **232**, 431 (EEP)
Efstathiou, G. 1996, in *Cosmology and large scale structure*, Les Houches, Session LX, Schaeffer, R. et al. (eds.), North-Holland
Ellis, R. S. 1997, *Annu. Rev. Astron. Astrophys.*, **35**, 389
Ferguson, H. C., & Babul, A. 1998, *Mon. Not. R. Astron. Soc.*, **296**, 585
Heyl, J., Colless, M., Ellis, R. S., & Broadhurst, T. J. 1997, *Mon. Not. R. Astron. Soc.*, **285**, 613
Hogg, D. W., & Phinney, E. S. 1997, *Astrophys. J*, **488** L95
Koo, D. C., & Kron, R.G. 1992, *Annu. Rev. Astron. Astrophys.*, **30**, 613
Kullback, S., & Leibler, R. A. 1951, *Ann. Math. Statist.*, **22**, 79
Lin, H., Kirshner, R. P., Shectman, S. A., Landy, S.D., Oemler, A., Tucker, D. L., & Schechter, P. L. 1996, *Astrophys. J*, **464**, 60
Lynden-Bell, D. 1971, *Mon. Not. R. Astron. Soc.*, **155**, 95
Marshall, H. L., Avni, Y., Tananbaum, H., Zamorani, G. 1983, *Astrophys. J*, **269**, 35
Rao, C. R. 1965, *Linear Statistical Inference and its Applications*, John Wiley & Sons, New York
Sakamoto, Y., Ishiguro, M., & Kitagawa, G. 1986, *Akaike's Information Criterion Statistics*, Kluwer Academic Publishers, Dordrecht:Reidel
Sandage, A., Tamman, G. A., & Yahil, A. 1979, *Astrophys. J*, **232**, 352 (STY)
Saunders, W., Rowan-Robinson, M., Lawrence, A., Efstathiou, G., Kaiser, N., Ellis, R. S., & Frenk, C. S. 1990, *Mon. Not. R. Astron. Soc.*, **242**, 318
Schechter, P. L. 1976, *Astrophys. J*, **203**, 297
Schmidt, M. 1968, *Astrophys. J*, **151**, 393
Shiryaev, A. N. 1996, *Probability*, 2nd edition, Springer Verlag, New York Inc.
Strauss, M. A., & Willick, J. A. 1995, *Phys. Rep.*, **261**, 271
Stuart, A., Ord, J. K., & Arnold, S. 1999, *Kendall's Advanced Theory of Statistics*, Vol. 2A, Classical Inference and Linear Model, 6th edition, Arnold, London